Java API for MIAPE Generation

Emilio Salazar Donate, Miguel Ángel López García, Salvador Martínez-Bartolomé, J. Alberto Medina-Aunon, Juan Pablo Albar

ProteoRed, Proteomics Facility, Centro Nacional de Biotecnología (CNB), Consejo Superior de Investigaciones Científicas (CSIC), E28049, Madrid, Spain.

The HUPO- PSI (Human Proteome Organization – Proteomics Standard Initiative) aims to define community standards for data representation in proteomics to facilitate data comparison, exchange and verification [1].

One of these standards is the MIAPE-specifications [2]. MIAPE, the Minimum Information About a Proteomics Experiment, contains the basic reporting guidelines for proteomics. Although, these guidelines are an invaluable support for experiments due to its acceptance by the proteomics researchers and journals, the level of fulfillment for every experiment requires a time consuming duty from the user side.

The **Miape JAVA API** (Application Programming Interface) developed by the CNB-ProteoRed Proteomics Bioinformatics Support group provides the user with a powerful, platform-independent programmatic interface to store, retrieve and export MIAPE documents in different formats. The libraries and the documentation are freely available at www.proteored.org/miape-api

The language of the API is **Java**, an object-oriented language, massively used in professional and academia environments due to its maintainability and flexibility, as well as its platform independence.

The aim of the API is to be integrated into third party software to automatically generate a MIAPE, decreasing the amount of extra work which this normally involves. So far, the user must manually submit this information independently, using the semi-automatic MIAPE generator tool [3], which allows the user to create a MIAPE, using templates with some initial information.

The API is divided in several independent modules, which can be extended for customization, (see Figure 1). The interaction between the modules is via the abstract entity Miape, which includes the information available about the experiment, regardless if it has been generate via XML, manually or it has been retrieved from the database.

Moreover, the API allows the user to convert the data from one module to another, favoring the exchange of information in very different formats.

The modules are:

- The **Database Manager** module: retrieves/stores documents from the ProteoRed Database, a relational database which has, so far, more than 700 documents. However, this module can be easily extended to be used in other Databases (relational or not).
- The **XML** module is able to parse files from different XML formats to generate the MIAPE documents.
- The **Factory** module creates a document manually by programmatically adding the information to the document.

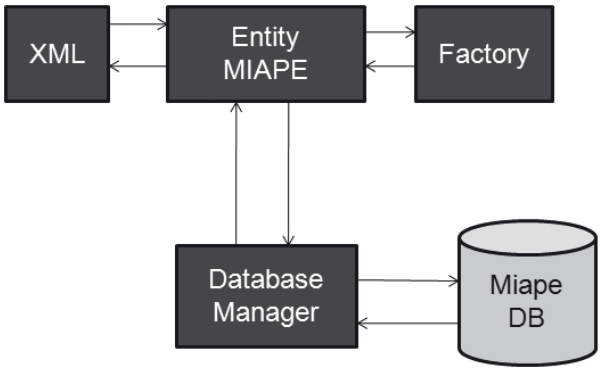

Figure 1. *The different modules of the Miape JAVA API and their interactions.*

References

[1] Kaiser J. Proteomics. Public-private group maps out initiatives. Science 2002;296:827.

[2] Taylor CF, Paton NW, Lilley KS, Binz PA, Julian RK, Jr., Jones AR, et al. The minimum information about a proteomics experiment (MIAPE). Nat Biotechnol. 2007; 25:887-93.

[3] Martínez-Bartolomé S, Medina-Aunon JA, Jones AR, Albar JP. "Semi-automatic tool to describe, store and compare proteomics experiments based on MIAPE compliant reports". Proteomics (in press) 2009.

The Proteomics Identifications database (PRIDE), its associated tools and the ProteomeXchange consortium

Juan Antonio Vizcaíno, Florian Reisinger, Richard Côté, Henning Hermjakob

EMBL-EBI, Wellcome Trust Genome Campus, Hinxton, Cambridge, United Kingdom

Abstract

The Proteomics Identifications Database (PRIDE, http://www.ebi.ac.uk/pride) has become one of the main repositories of mass spectrometry derived proteomics data. In this communication we will summarize the main capabilities of the PRIDE system, including its associated tools. Finally, we will introduce the ProteomeXchange consortium, as a collaborative approach to share proteomics data between the most important proteomics repositories.

The PRIDE Proteomics Identifications database (http://www.ebi.ac.uk/pride) at the European Bioinformatics Institute (EBI) provides users with the ability to explore and compare mass spectrometry (MS) based proteomics experiments that reveal details of the protein expression found in a broad range of taxonomic groups, tissues and disease states [1]. PRIDE stores three different kinds of information: peptide and protein identifications derived from MS or MS/MS experiments, MS and MS/MS mass spectra as peak lists, and any and all associated metadata. PRIDE is now the recommended submission point for proteomics data for several journals such as *Nature Biotechnology, Nature Methods, Molecular and Cellular Proteomics*, and *Proteomics*.

1. PRIDE associated tools

PRIDE relies heavily on two additional tools: the Ontology Lookup Service [2] (OLS, http://www.ebi.ac.uk/ols), and the Protein Identifier Cross-Referencing system [3] (PICR, http://www.ebi.ac.uk/Tools/picr). OLS provides convenient and powerful access to a large number of biomedical ontologies and controlled vocabularies (CVs). PRIDE takes advantage of OLS to store, structure, and present any and all metadata annotations on experiments, proteins, peptides and mass spectra.

The PICR tool on the other hand, is built to overcome one of the most recurrent problems in proteomics: the existence of heterogeneous and changing identifiers or accession numbers referring to the same protein in different databases. PICR is used to map all the submitted protein identifications in PRIDE to all known accession numbers for those proteins in the most important protein databases (including UniProt, IPI, Ensembl and RefSeq, among others). Therefore, protein identifications in PRIDE that were originally derived from different databases, or from different time points of the same database, thus become fully comparable. In addition to these two established tools, a new application called Database on Demand [4] (DoD, http://www.ebi.ac.uk/pride/dod) has recently been added to the PRIDE toolkit. This tool allows custom sequence databases to be built in order to optimize the results from search engines for gel-free proteomics experiments.

2. ProteomeXchange Consortium

One of the reasons why proteomics data sharing is not a universal fact yet is the heterogeneity of the